DOCUMENT RESUME

ED 322 164 TM 015 278

AUTHOR Friedman, Charles B.; Ho, Kevin T.
TITLE Interjudge Consensus and Intrajudge Consistency: Is It Possible To Have Both in Standard Setting?
PUB DATE Apr 90
NOTE 19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Evaluators; Feedback; *Interrater Reliability; *Licensing Examinations (Professions); Medicine; Minimum Competency Testing; Multiple Choice Tests; *Standardized Tests; Test Construction
IDENTIFIERS Angoff Methods; *Consensus; *Standard Setting

ABSTRACT

Eleven judges representing 11 different geographic regions in the United States participated in a standard-setting session designed to determine the possibility of obtaining interjudge consensus and intrajudge consistency simultaneously. Each judge had experience in the field for which standards were being set. The judges rated 65 multiple-choice test items that are part of a national health certification examination. All items had been field tested. The judges participated in a rater-training session that included a discussion of the Angoff method and the conceptualization of skills needed for minimum competence. After all items were rated, handouts containing a list of all ratings for all judges were reviewed and discussed by the judges. All judges changed their ratings on at least 4 of the items, and 2 judges changed their ratings on 20 or more items. Results indicate that the procedures used to improve interjudge consensus (that is, provision of feedback via the handouts) did not have an adverse affect on intrajudge consistency. In fact, results show that techniques designed to improve consensus also improved consistency. Two data tables and two bar graphs are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kevin T. Ho

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED322164

Interjudge Consensus and Intrajudge Consistency: Is It Possible To Have Both in Standard Setting?

Charles B. Friedman

Kevin T. Ho

The Psychological Corporation

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, 1990.

TM 015278

BEST COPY AVAILABLE

Interjudge Consensus and Intrajudge Consistency: Is it Possible to Have Both in Standard Setting?

As the demand for licensing, certification, and the establishment of minimal standards of competency has increased, so has the attention to the processes by which standards are established. Much effort has been expended to understand and improve the standard-setting process. The initial and continuing focus has been on judgmental methods and procedures, e.g., Nedelsky (Nedelsky, 1954) and Angoff (Angoff, 1971).

More recently, measurement experts have been concerned with factors affecting the quality (validity) of the content experts' judgments and with the means of improving the accuracy of the those judgments. Modifications have been suggested to improve both the consensus between judges in rating items as well as the reliability or consistency of an individual judge's ratings.

These modifications have included the use of empirical item data (Smith and Smith, 1988), the comparison and discussion of judges' ratings during the rating session (feedback), the provision for judges to change their ratings after the discussion session, and the trimming of judges' scores prior to calculating a final standard.

Underlying judgmental methods in standard setting is the assumption that judges will reach agreement in their understanding of minimal competence and, in effect, consensus in judging items. Juxtaposed with the interjudge consensus, however, is the notion of intrajudge consistency, or the degree to which an individual judge's rating of a set of items is consistent with itself and with the performance on those items by the minimally competent candidate whom that judge has in mind (van der Linden, 1982; Plake, Melican, & Mills, 1989).

These two notions, however, may not be complementary and procedures designed to increase interjudge consensus may in fact decrease intrajudge consistency. For example, some judges may be easily swayed in their item estimates by a highly vocal individual. In such a case, an attempt to increase consensus may cause the judge to overreact, to make changes thoughtlessly, and decrease consistency.

Studies have generally defined consistency only as interjudge consistency or consensus and have either ignored intrajudge consistency or viewed it as the correlations of the individual judge's ratings with actual p-values for the items being rated (Smith and Smith, 1988). In some cases, internal reliability estimates were used as indicators of intrajudge consistency and the accuracy of the standard-setting process (Melican, Mills and Plake, 1988; Garrido and Payne, 1987).

The application of an Item Response Theory (IRT) model, as suggested by van der Linden (1982) provides an alternative method for viewing intrajudge consistency. This model suggests using the judge's conceptualization of a minimally competent candidate to set a standard by which that judge's internal consistency can be measured. The individual judge's ratings of all the items are used to obtain an estimate of that judge's conceptualization of the ability level of the minimally competent candidate. That ability level is then used in conjunction with actual item parameters to compute a standard for comparison.

This study proceeds from the assumption that both interjudge consensus and intrajudge consistency are important indicators of the accuracy of the standard-setting process. For the purpose of this paper the term consensus will be used for the agreement between judges (interjudge) and the term consistency will be used for the agreement of an individual judge's ratings (intrajudge) and his conceptualization of the minimally competent candidate. It seeks to examine the potential interaction between interjudge consensus and intrajudge consistency.

The following questions are addressed:

1. Does providing judges with an opportunity to modify their initial item ratings affect the intrajudge consistency of the judges' ratings?
2. Can measures of intrajudge consistency be used to eliminate judges with low measures from the rating pool? What effect would this have on the standard-setting score?

This study addresses the general question, "Is it possible to have both interjudge consensus and intrajudge consistency in the judges' ratings of items?"

Methods

Participants

Eleven judges representing eleven different geographic regions in the United States participated in the standard-setting session. Each judge had experience in the field for which the standard was being set.

Instrument

The eleven judges rated 65 multiple-choice test items that are part of a national health certification exam. All of the items had been field tested. The results of those judges' ratings serve as the data for this study.

Rating Procedure

The eleven judges participated in a rater-training session which included a discussion of the Angoff method and the conceptualization of the skills needed for minimum competence.

As part of a practice rating exercise, each judge was given a set of items to rate. Each judge then compared his or her ratings to both the ratings of the other judges and also to the relative difficulty of an item as expressed by its p-value from the results of the field testing. After the training session, the judges were asked to rate the real items.

Each judge was provided with a rating sheet which consisted of item identifiers and a list of probabilities. The probabilities ranged from .10 to .90 in increments of .10. Each judge then circled his or her choice as the most appropriate probability for each item. This constituted the first pass through the items.

After each judge had rated all the items, they were given a short break during which a handout was prepared containing a list of all the ratings for all the judges. The handouts were distributed to all judges. Each judge was then able to compare his or her rating of an item with the ratings of that item by other judges. They were free to discuss any items which lacked consensus. After these discussions, the judges were allowed to change their item ratings. This constituted the second pass through the items.

All eleven judges changed their ratings on at least four of the items and two judges changed their ratings on twenty or more of the items.

Analysis

The analyses in this study had two purposes: (a) to compute indices of intrajudge consistency for all judges; and (b) to apply those measures of consistency to evaluate the standard-setting process. In other words, measures of consistency were first computed for each judge on both the first and second pass through the items. Those measures were then used to compare judges' ratings in the first and second pass.

Indices of consistency were computed using a method proposed by van der Linden (1982). This technique uses the judge's ratings of items in

conjunction with IRT item parameters to obtain probabilities of correct responses for each item on the exam by the minimally competent candidate. Since these probabilities for a correct response are based on the ability level (or theta) for that judge's minimally competent candidate, the resulting IRT derived probabilities can be used to measure that judge's subjective ratings of each item.

There are two measures of consistency produced: the average absolute error of specification (E) and an index of consistency (C_1). E represents the amount of error that the judge had in rating the items. In other words, the difference between that judge's ratings and the probabilities of a correct response by the minimally competent candidate conceptualized by that judge. The larger the value of E, the greater the rating error.

The value of E is used in computing the index of consistency. The index of consistency (C_1) is a measure of the rating consistency of that judge with his or her personal concept of the minimally competent candidate.

The value of C_1 can range from 0 to 1. The larger the value of C_1, the more consistent the judge's ratings. It is important to note that E and C_1 are inversely related--as the value of E increases, the value of C_1 decreases. In other words, as a judge's error of specification increases, that judge's consistency decreases.

The following procedures were applied to compute E and C_1. First, a one-parameter Item Response Theory model was used to compute item difficulties for the 65 test items based on actual examinee data. Along with item parameters and statistics, a true-score-to-theta table was also generated. This table provided the conversion of a candidate's true score to an ability estimate.

The following steps were then applied for each judge:

1. The judge's ratings for all items were summed. This not only represented that judge's passing standard, but also served to represent the estimate of the minimally competent candidate's true score (van der Linden, 1982).

2. The candidate's true score was converted to a theta value by using the true-score-to-theta table generated as part of the item parameter estimation procedure.

3. The theta value was then used in the following formula to compute the probabilities of the candidate passing each of the 65 items.

$$P_i = [\ 1 + \exp^{-(\text{theta} - b)}\ ]^{-1}$$

where:

P_i = the probability of a correct response for item i and b = the item difficulty for item i.

4. The index of consistency (C_1) is computed for this judge by using the following formula:

$$C_1 = \frac{M - E}{M}$$

where:

$$M = \sum e/n$$

where:

n = number of items

e = the maximum of (p_i, 1 - p_i) [e=the larger of the two values]

and:

$$E = \sum |p^s - p_i| / n$$

where:

p^s = the judge's rating on item i

p_i = the probability of success on item i as computed in step 3 above

n = number of items

The second part of the analysis dealt with using the measures of consistency in evaluating the standard-setting process. As was stated earlier, the judges were allowed to change their initial item ratings after being provided feedback in the form of a report of the item ratings for all judges in the group. This feedback/modification cycle is a technique used to increase consensus.

Indices of E and C_1 were generated on the ratings of all judges for the first and second passes through the items. A comparison of the mean E and C_1 for all judges for the first and second passes will show if consistency increased or decreased on the second pass. The index of consistency for individual judges on the first and second passes will also be compared.

Two methods for eliminating judges were examined in order to obtain an adjusted cut score. The first method considered consensus as the primary factor; therefore, this method eliminated the judge with the highest sum of item ratings and the judge with the lowest sum of item ratings (trimmed mean or outlier method). The second method considered consistency as the primary factor; therefore, this method eliminated the two judges with the lowest consistency scores (C_1). The mean of the sums of item ratings for the remaining judges for both methods were then compared.

Results

The results of the analyses of the first and second passes for all eleven judges are found in Table 1. The following information is presented for each judge's first and second ratings; the sum of the probabilities for all items (the "cut score"), the average absolute error

for all items (E) and that judge's measure of consistency (C_1). In addition, the difference between each judge's first-pass C_1 and second-pass C_1 are presented. The number of item ratings that each judge changed from the first to the second pass is also noted.

For example, in looking at Table 1 it can be seen that the sum of ratings for judge 1 in the first pass was 36.70, the error of specification (E) was .192, and the measure of consistency (C_1) was .728. In the second pass, judge 1 changed the rating of 12 items and had a sum of ratings of 38.40, an E of .178, and a C_1 of .75. Furthermore, when the sum of judge 1's scores increased on the second pass, his or her intrajudge consistency also increased by .022. The numbers for the other judges are similarly displayed.

As indicated in Table 1, the fewest number of ratings changed was four and the greatest number changed was twenty, with a mean of 12.82.

Insert Table 1 here

The sums of the judges' ratings on the first pass ranged from 36.50 to 45.30 with a mean of 39.79 and a standard deviation of 3.25. The range on the second pass was from 38.40 to 46.00 with a mean of 41.78 and a standard deviation of 2.40. Ten of the eleven judges raised the cut score on the second pass (see Figure 1); thus, the cut score on pass two is higher than on pass one and the standard deviation on pass two is lower than on pass one.

Insert Figure 1 here

The mean error (E) for all judges decreased from .183 in the first pass to .165 in the second pass. Conversely, the mean of the index of consistency (C_1) increased from .745 in the first pass to .775 in the second pass. Note that the index of consistency (C_1) for all judges increased from pass one to pass two by an average of .029. It is essential to note that in no case did the index of consistency decrease for any judge between pass one and pass two (see Figure 2).

Insert Figure 2 here

Table 2 compares two methods of eliminating judges from the pool after the second pass. From a consensus perspective, the outlier method was used to eliminate two judges. The judges are ranked according to the sum of their item ratings (smallest to largest), and the two outliers were eliminated. In the consistency method, the judges are ranked by their C_1 index (lowest to highest), and the two judges with the lowest consistency were eliminated. Each judge's rank, the sum of ratings, and the decision to keep or eliminate that judge is presented for both the outlier and the lowest consistency methods. Additionally, the value of C_1 is presented for the consistency method.

It can be seen that for judge 1, the rank in the outlier method was 1, the sum was 38.40, and the decision was made to eliminate this judge. In the consistency method, the same judge had a rank of 2, a C_1 value of .75, a sum of 38.40, and the decision was made to also eliminate this judge. The results for the other judges were similarly reported.

The outlier method eliminated the two judges with the highest and lowest sum of item ratings. These were judges 4 and 1 respectively. The consistency method eliminated the two judges, judges 1 and 11, with the

lowest C_1 ratings. The adjusted mean sum of item ratings for the outlier method was 41.69, while the adjusted mean sum of item ratings for the consistency method was 42.10.

Insert Table 2

It is interesting to note that judge 1 would have been eliminated after the second pass using either method. Judge 1's concept of a minimally competent candidate was not only lower than the other judges, but judge 1 was also inconsistent when specifying the probablilites of success for that candidate on each of the items. It is also important to note that, while the outlier method would have eliminated judge 4, examination of the C_1 value for judge 4 shows that judge 4 had the highest consistency rating of all the judges. This means that while judge 4's concept of a minimally competent candidate was higher than any other judge's, judge 4 was most consistent in specifying the probabilities of that candidate passing each of the items.

Discussion

The purpose of this study was to see if it is possible to have both consensus and consistency in standard setting. The procedures used to improve interjudge consensus (providing feedback) did not have an adverse affect on intrajudge consistency; in fact, the results showed that techniques designed to improve consensus also improved consistency. It may not be unusual for a consensus technique to improve consistency. During the first pass, judges may drift away from their conceptualization of the minimally competent candidate; during the second pass with the

discussion, judges are permitted to fine tune their internal conceptualization and to increase their internal consistency.

Using measures of consistency as criteria for eliminating judges in the final setting of the cut score was also explored. As reported, those judges who were eliminated based on their intrajudge consistency scores were not always the same judges who were eliminated by the outlier method. Although one judge would have been eliminated by either method, the outlier method would have also eliminated the judge with the highest index of consistency. This suggests the importance of using both measures of consensus and consistency in the standard-setting process.

Although indices of consistency were generated and used in the study to eliminate judges in setting a standard, we need to examine the impact of those decisions on the actual passing standard. It should be pointed out that although the two methods for adjusting the cut score eliminated different judges, the cut score for the groups selected by the two methods did not change. The cut score for the group selected by the outlier method (41.69) was not much different than the cut score for the group selected by the lowest consistency method (42.10). This fact and the more general problem of how to address conflicts between measures of consensus and consistency point to the need to further investigate the proposed index of intrajudge consistency and its place in the standard-setting process.

There is currently no measure to gauge the relative strength of the index of intrajudge consistency and its impact on the cut score. Until more is known about the index, it cannot be used to its full potential. Given a different group of judges in a different situation, the values of C_1 and its relationship to the standard-setting process might be different than the results found here. Judgment panels, by their nature,

are subjective. The above suggests that a measure of consistency can be used as a criterion to compare the results of different panels. If two standard-setting panels are convened and the recommended cut scores are different, the internal consistency indices could be used in selecting which score to use.

Some standard or measure of the relative strength of the index of intrajudge consistency needs to be developed. For what values of C_1 should judges be dropped from any further analysis? At what point, if any, does the index of consistency outweigh any other measures of interjudge consensus? What is a good overall C_1 for the entire group?

The use of an IRT-based measure of intrajudge consistency is a valuable tool, because it is the only measure that can generate empirical probabilities for the minimally competent candidate as envisioned by the judge. Because these empirical probabilities are based on the judge's description of the minimally competent candidate and on actual item information, this model provides us with a means of judging each judge against an appropriate standard, i.e., his or her own image of the minimally competent candidate, rather than the ratings of other judges or the data based on a group of examinees.

Although the initial uses of the IRT-based index of intrajudge consistency have yielded some valuable and interesting data, the true extent of its usefulness will not be known until it is subjected to additional examination. The use of an intrajudge-consistency measure with other real data or in a Monte Carlo simulation may provide a clearer interpretation of a consistency index as well as the information needed to better evaluate the interaction of consensus and consistency in the standard-setting process.

References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), Educational measurement. Washington, DC: American Council on Education.

Garrido, M. & Payne, D. (1987). An experimental study of the effect of judges' knowledge of item data on two forms of the Angoff standard setting method. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C.

Melican, G.J., Mills, C.N., & Plake, B.S. (1988). Accuracy of option elimination prediction: Effect of a technique to improve rater agreement with empirical data. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Plake, B.S.; Melican, G.J., and Mills, C.N. (1989). Factors influencing the degree of intrajudge consistency during the standard setting process. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Smith, R.L. and Smith, J.K. (1988). Differential use of item information by judges using Angoff and Nedelsky Procedures. Journal of Educational Measurement, 25, 259-274.

van der Linden, W.J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 19, 295-308.

Table 1.

Results of Judges' First Ratings and Second Ratings.

	First Pass			Second Pass				
Judge	Sum[a]	Error	C_1	Sum	Error	C_1	Difference[b]	#Changes
1	36.70	.192	.728	38.40	.178	.750	.022	12
2	41.50	.191	.739	42.50	.167	.774	.035	13
3	37.60	.178	.749	42.10	.157	.785	.036	21
4	44.90	.168	.776	46.00	.153	.798	.022	12
5	36.50	.186	.736	39.90	.156	.782	.046	20
6	39.80	.165	.770	42.60	.156	.789	.019	13
7	39.10	.182	.745	41.50	.166	.773	.028	12
8	45.30	.168	.776	45.10	.157	.790	.014	4
9	37.70	.193	.728	40.30	.163	.773	.045	17
10	36.50	.186	.737	38.70	.165	.769	.032	13
11	42.10	.209	.715	42.50	.194	.737	.022	4
mean	39.79	.183	.745	41.78	.165	.775	.029	12.82
s.d	3.25			2.40				

[a]sum of judges' ratings for all items
[b]difference = C_1 of Pass 2 - C_1 of Pass 1

Table 2.

Elimination of Judges After Pass 2.

	Outlier Method			Consistency Method			
Judge	Rank[a]	Sum	Decision	Rank[b]	C_1	Sum	Decision
1	1	38.40	Eliminate	2	.750	38.40	Eliminate
2	7	42.50	Keep	6	.774	42.50	Keep
3	6	42.10	Keep	8	.785	42.10	Keep
4	11	46.00	Eliminate	11	.798	46.00	Keep
5	3	39.90	Keep	7	.782	39.90	Keep
6	9	42.60	Keep	9	.789	42.60	Keep
7	5	41.50	Keep	5	.773	41.50	Keep
8	10	45.10	Keep	10	.790	45.10	Keep
9	4	40.30	Keep	4	.773	40.30	Keep
10	2	38.70	Keep	3	.769	38.70	Keep
11	8	42.50	Keep	1	.737	42.50	Eliminate
adj. mean		41.69[c]				42.10[d]	

[a]Ranked according to sum of ratings; - outliers are highest and lowest value

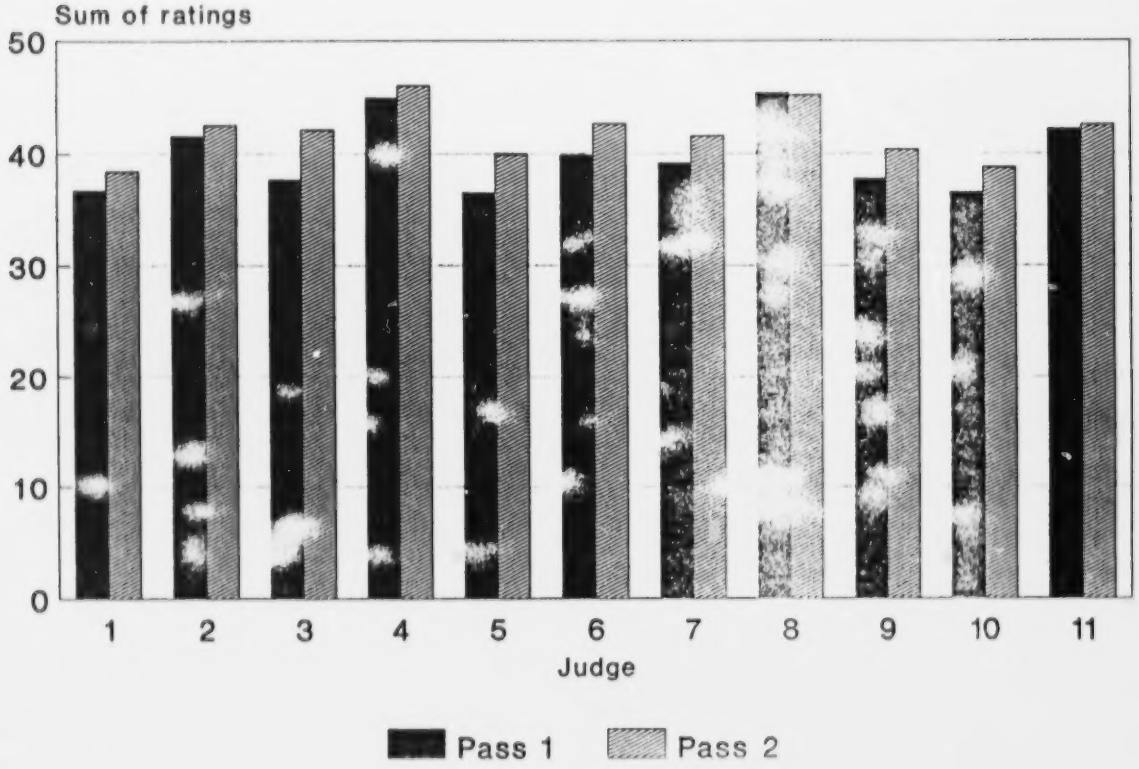
[b]Ranked according to measure of consistency - two lowest C_1 values are dropped (11 is most consistent and 1 is least)

[c]Mean of all judges except 1 and 4

[d]Mean of all judges except 1 and 11

9963I

Figure 1. Sums of ratings-first pass vs. second pass.



sum = sum of ratings for all items

Figure 2. Index of consistency-first pass vs. second pass.